Application of Normal Multivariate Binary Scale Mixing in Regression Model

Yinshan Jiang

South China Business College (Scbc) of Guangdong University of Foreign Studies (Gdufs), Guangzhou, 510545, China

Keywords: Binary Scale of Positive Polymorphic Variables; Logistic Regression Model; Credit Risk; Factor Analysis

Abstract: Credit risk is the most important risk faced by commercial banks, and it has a crucial impact on the bank's continued sound operation. How to effectively manage credit risk on the basis of accurate measurement is a very challenging issue facing commercial banks. This article uses public financial data of listed companies to establish a Cox credit risk normal multivariate binary scale mixed regression model. The selected indicators have high dimensionality and strong correlation, and contain a lot of redundant information. How to select the covariates that really affect the response normal multivariate from the much information for modeling is very necessary. In this paper, by analyzing the application scope, data requirements, prediction accuracy and stability of each model, this paper finds that the Logistic regression model is more suitable for the actual situation in China in terms of input data and premise assumptions. Therefore, this paper finally chose the Logistic regression method to establish credit Risk measurement model, and an empirical test of the normal multivariate binary scale for the effect of the regression model. When using the logistic regression method to establish the credit risk measurement, this paper proposes a method of objectively selecting the input parameters of the logistic model. The experimental results show that the normal multivariate binary scale is used to distinguish the normal companies above 80%. However, there is still a significant gap compared with the logistic model's judgment accuracy of training samples above 90%.

1. Introduction

The domestic academic community has conducted a large number of normal multivariate binary scales theoretical and empirical studies on establishing a stable and effective credit risk measurement regression model suitable for the actual situation of China's commercial banks, and has achieved many positive research results. However, in the selection of research indicators, default there is still some defects in the determination of the ratio of the sample and the non-default sample [1]. In view of the above deficiencies in the existing research, this article is prepared to start research from the above two aspects, and explore to establish a normal multivariate binary scale credit risk measurement regression model that is suitable for the actual situation of China's commercial banks [2-3]. Therefore, whether this article is to establish an effective internal normal multivariate binary scale credit risk measurement regression model for the banking industry or to conduct in-depth research on this issue by later scholars, it has certain reference and guidance significance [4].

Based on the above-mentioned problems, the research idea of this paper is: first, through a qualitative summary of the concept and characteristics of credit risk, and then to summarize and sort out various measures and management models of credit risk [5], and explore their respective Advantages, disadvantages and applicability. Finally, logistic regression is used to establish a credit risk measurement model based on the objective selection of the normal multivariate binary scale for the input indicators, and the effect of the model is empirically tested and the Fisher discriminant model is established with traditional methods. Compare [6].

In this paper, the method of combining normative analysis and empirical research is used in the research to strive for a comprehensive analysis of the research object and draw meaningful conclusions for practical application and subsequent research. This article uses normative analysis

to sort out and introduce the credit risk measurement regression model, expounds the connotation, characteristics, impact and management methods of credit risk, and studies the regression model of normal multivariable binary scale credit risk for the following Made conceptual bedding. Use the sample data of listed companies to build a logistic credit risk measurement regression model. A stepwise discriminant method was used to establish the Fisher multivariate discriminant model for comparison with the Logistic regression model.

2. Proposed Method

2.1 Selection of Normal Multivariate Binary Scale for Regression Models

(1) Based on stepwise regression method

Stepwise regression is a traditional normal multivariate selection method [7]. The basic idea is: according to the pre-set significance standard, the normal multivariable is introduced into the model one by one according to the significance size, from large to small, and the non significant normal multivariable is eliminated by Wald test in each step. Here, the Wald statistic is:

$$Wald_i = \frac{\hat{\beta}_i}{S_{\beta_i}} \tag{1}$$

Where $\hat{\beta}_i$ represents the maximum likelihood estimate of the ith normal multivariable and S_{β_i} represents the standard error. The Wald statistic obeys the chi square distribution of degree of freedom 1. Suppose S_{β_i}= 0, first calculate the p value of each normal multivariable, and then compare it with the given significance level. If the normal multivariable is significant, keep it, if not, delete it. Repeat the above steps until the normal multivariable in all models is significant and the normal multivariable out of all models is not significant [8].

(2) Lasso based method

In recent years, there are a lot of research results about regression model. However, when the dimension of normal multivariable is large, the traditional regression model is not very ideal for the estimation of model coefficients, and when there is a collinearity, it may also lead to a large deviation of the results. At this time, data need to be dimensionally reduced and collinearity processed, that is to say, normal multivariable selection [9]. The method of applying lasso penalty function to regression model is very similar to that of linear model, as long as the negative log likelihood function of regression model is used to replace the sum of squares of residuals in the equation of linear model. In general, the maximum likelihood estimator β can be obtained by maximizing the partial partial partial log likelihood function $l_n(\beta)$ of the regression model. In this paper, we first minimize the opposite number of the logarithmic likelihood function of the formula, then $\hat{\beta} = argmin - l_n(\beta)$ at this time, together with lasso penalty function, is expressed in the form of unconstrained conditions. Therefore, the lasso estimator under the regression model can be defined as:

$$\hat{\beta}=\text{argmin}\{\frac{1}{n}\sum_{i=1}^{n}\{-\beta^T Z_i + \ln(\sum_{j\in R_i}\exp(\beta^T Z_j))\} + \lambda\sum_{j=1}^{k}\beta_j \tag{2}$$

For lasso penalty function, this paper uses the least angle regression algorithm (LARS) to solve. Lars algorithm is proposed for lasso method. Compared with other algorithms, Lars algorithm can effectively solve lasso problem. Lars algorithm needs K steps to find the regularization path of L1 norm (where k is the number of normal multi variables). The solution of this algorithm is to continuously fit the residual, and gradually reduce the corresponding residual by adding new normal multi variables. The solution path of Lars algorithm meets the following requirements: the correlation coefficient of the selected normal multivariable is equal to the correlation coefficient of the current residual. The specific methods are as follows:

First, let all the normal multivariable coefficients β_j be equal to zero, find a co normal multivariable with the largest correlation with the response normal multivariable y, and record it as

z_1, and add z_1 to the normal multivariable set ψ; secondly, continue to move forward on the solution path of the normal multivariable $z_1$according to the same method until a new normal multivariable z_2 is found, making the normal multivariable z_1, The correlation coefficient of z_2 is equal to the correlation coefficient of the current residual, and then z_2 is added to the normal multivariable set ψ; thirdly, take the isometric of z_1 and z_2 as the direction of motion, find a new solution path along this direction, and calculate the corresponding residual until z_3 is found, Make the normal multivariable z_3 and residual R have the same correlation, and then add them to the normal multivariable set ψ; and so on until all the normal multivariable are in ψ.

2.2 Binary Logistic Regression Model in Letter

(1) Logistic regression model

Logistic model is a statistical method for non-linear classification, using binary logistic probability function as the model equation. Logic functions are also called development functions. The binary logistic regression model is mainly applicable to the case where there are only two cases of variable values. For example, when providing the default or non-default status of a corporate loan, a Logistic regression model can be used to determine. The Logisti regression model not only classifies companies into default and non-default categories based on the designated division points, but also returns the expected default probability of each loan company. In the research process of this article, the value of the dependent variable is 0 and 1, when the value of the dependent variable is 1, it means that the company is the default value, and the value is. Indicates that the company is not the default company. The formula of the logistic regression model is as follows:

$$P_i = \frac{1}{1+e^{-z_i}} \tag{3}$$

$$Z_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ki} \tag{4}$$

Z_i is the joint action of all independent variables (here refers to the financial ratio index), which is called the score of the financial status of the ith loan company; P_i is the probability when the dependent variable is taken as 1, which is the default probability of the loan company; x_{ki} is the K financial ratio index variable of the ith loan company in the credit risk assessment; $\beta_k$is the regression coefficient of. According to the definition of default, it can be seen that the greater the default, the greater the possibility of enterprise default.

Z_i is the common result of independent variable (here is financial ratio), which is called the financial statement of the ith loan company. P_i is the probability of dependent variable 1, that is, the probability of default of loan company, and x_{ki} is the credit risk assessment. The k-th variable ratio β_k of the i-th loan company is the regression coefficient of x_{ki}. According to the definition of default, it can be seen that the higher the default, the greater the opportunity for the company to charge.

(2) Correlation analysis among factors

Because there is a strong correlation between the financial ratios of enterprises, multi culture will affect logistic. In this paper, factor analysis is used to solve the influence of multi centrality between variables, because factor analysis can create new common factors through the internal correlation of variables, so as to reduce the dimension of variables. In order to overcome the multi centrality between independent variables, in addition, the main information of the original indicators can be retained in the coefficient analysis process, and each indicator is standardized, which makes the scale impact of each economic indicator comparable.

(3) Back testing of training samples

As a credit risk measurement model, logistic model can divide the input samples into two categories: non payment company and non payment company. That is, the model can compare the expected default probability of each enterprise with the predetermined dividing point to compare the default probability. Businesses larger than the critical point are classified as unfair companies, and businesses with a probability of default less than the critical point are classified as unfair companies. The model can use two types of errors to classify businesses by fragmentation points. These two types of errors are called first type errors and second type errors respectively. The first type of error refers to the wrong judgment of non defaulting company as default company; the

second type of error refers to the wrong judgment of non listed company as default company. In the credit risk analysis, two types of errors will lead to the loss of commercial banks, but because of the characteristics of loans, the loss caused by the two types of errors are very different.

3. Experiments

3.1Fishe: Discriminant Analysis Method

Fisher discriminant analysis is a multivariate statistical method to discriminate the type of sample. The multivariate discriminant analysis model is mainly based on the classification of some known research objects, and based on these research objects, a discrimination criterion is established to determine which category the new sample belongs to. For the first time, the multivariate discriminant model was used to judge the default situation of an enterprise, and the famous Z-Score five-variable model was established. Since then, the multivariate discriminant model has been widely used in the field of credit risk measurement. The theoretical and practical circles have used multivariate discriminant models. A lot of research has been done on the credit risk measurement of the model, and great results have been obtained. From the previous analysis of this paper, we can see that although the multivariate discriminant model has certain advantages in practical applications, it also has some problems in the model itself. In this section, Fisher's multiple discriminant model will be used to conduct an empirical study of the default situation of the company for comparison with the results of the Logistic model.

3.2 Sample Selection

Select the 110 ST companies used above as a sample of defaulting companies, and select 110 normal companies that are similar in size and industry to the above ST as a non-defaulting sample from a total of 1,260 normal companies in a 1: 1 matching ratio, a total of 220 Listed companies serve as samples for empirical research. The above 220 listed companies are randomly divided into two groups of training samples and test samples according to a ratio of approximately 7: 3 for modeling and testing the model.

4. Discussion

4.1 Training Sample Back Analysis

The training samples used to establish the discriminant model are substituted into the established Fisher: multivariate discriminant model. The F_0 value and F_1 value of each sample company are calculated, and the two values are compared. If F_0 is greater than F_1, the listed company will be judged as ST company, i.e. default company. If F_0 is less than F_1, the listed company will be judged as normal company. The judgment results obtained through calculation are shown in Table 1 below:

Table 1. Classification results of training samples

		The current transaction status	Forecasting group members		Total
			0	1	
Initial	Count	0	64	14	78
		1	11	61	72
	%	0	82.1	17.9	100
		1	15.3	84.7	100

As shown in Table 1 above, it can be seen that the overall judgment accuracy of the multivariate discriminant model on the training samples reaches 83.3%. Among them, the model correctly identified the ST company as the correct identification rate of the ST company of 84.7%, and 15.3% of the ST companies were misjudged as normal companies; the discriminant model correctly identified the non-ST companies as 82.1%, and 17.9% Non-ST companies were misjudged as ST companies. The model's missed diagnosis rate for ST companies is 15.3%, and the false positive rate for non-ST companies is 17.9%. In general, the recognition effect of the model is good, and the

correct rate of discrimination for ST companies and normal companies is more than 80%. However, there is still a significant gap compared with the 90% accuracy rate of the training samples in the Logistic model.

4.2 Test Sample Back Analysis

In order to test the applicability of the model to companies outside the model sample, this article uses the test samples to substitute the established Flshe; a multivariate discriminant model, calculate the F.0 value and monthly value of each test sample company, and compare the two values. If it is greater than double, the listed company is judged to be an ST company, which is a default company. If it is less than double, the listed company is judged to be a normal company. The calculation result obtained by calculation is shown in Figure 1 below:

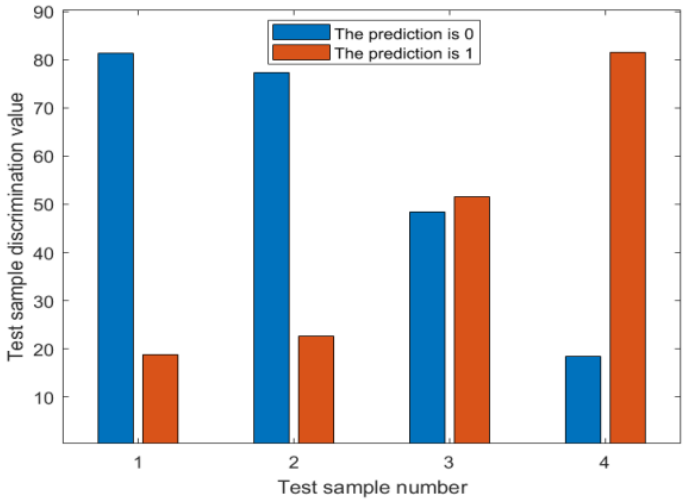

Figure 1. Discriminant analysis of test samples

As shown in Figure 1 above, it can be seen that the overall judgment accuracy of the multivariate discriminant model on the test samples outside the model reaches 81.4%. Among them, the model correctly identified ST companies as the number of ST companies as 31, the model's correct recognition rate was 81.58%, and 18.42% of ST companies were misjudged as normal companies; the model's correct recognition rate for non-ST companies It was 81.25%, and 18.75% of non-ST companies were misjudged as ST companies. The model's missed diagnosis rate for ST companies is 18.42%, and the false positive rate for non-ST companies is 18.75%. Judging from the test results of the test samples on the model, the model also has a good recognition effect on the new samples. The accuracy rate of ST companies and normal companies is more than 80%. The model established has strong stability.

Conclusions

Establish a gistic credit risk measurement model by using a full-sample modeling method instead of the 1: 1 ratio of ST companies and normal companies, and conduct training sample testing and test sample testing on the actual prediction effect of the model It can be found that both the training samples used for modeling and the test samples outside the modeling samples have higher prediction accuracy, and the overall judgment accuracy rates are 90.7% and 86.4%, which are close to 90%. The prediction accuracy rate of% can basically meet the needs of banks for measuring corporate credit risk.

References

[1]. Zellner, Arnold. Corrigendum: "Bayesian and non-Bayesian analysis of the regression model with multivariate Student- t error terms" (J. Amer. Statist. Assoc. 71 (1976), no. 354, 400–405). [J]. human brain mapping, 2015, 36(4):1292-1303.

[2]. Marcin Michalak, Emilia Gąsiorowska, Ewa Nowak-Markwitz. Diagnostic value of CA125, HE4, ROMA and logistic regression model in pelvic mass diagnostics – our experience [J]. Ginekologia Polska, 2015, 86(4):256-261.

[3]. Xiao Ma, Qiao Liu, Zhenyu He. Visual Tracking via Exemplar Regression Model [J]. Knowledge-Based Systems, 2016, 106(C):26-37.

[4]. Chagny, Gaëlle, Roche, Angelina. Adaptive estimation in the functional nonparametric regression model [J]. Journal of Multivariate Analysis, 2016, 146(2):105-118.

[5]. Bushra Naz Soomro, Liang Xiao, Lili Huang. Bilayer Elastic Net Regression Model for Supervised Spectral-Spatial Hyperspectral Image Classification [J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2017, 9(9):4102-4116.

[6]. Gergely Hegyi, Miklós Laczi. Using Full Models, Stepwise Regression and Model Selection in Ecological Data Sets: Monte Carlo Simulations [J]. Annales Zoologici Fennici, 2015, 52(5-6):257-279.

[7]. Manickavasagar Kayanan, Pushpakanthie Wijekoon. Performance of Existing Biased Estimators and the Respective Predictors in a Misspecified Linear Regression Model [J]. Open Journal of Statistics, 2017, 07(5):876-900.

[8]. Edwin M. M. Ortega, Gauss M. Cordeiro, Ana K. Campelo. A power series beta Weibull regression model for predicting breast carcinoma [J]. Statistics in Medicine, 2015, 34(8):1366-1388.

[9]. Dias, SÃ³nia, Brito, Paula. Linear regression model with histogramâ valued variables [J]. Statistical Analysis & Data Mining the Asa Data Science Journal, 2015, 8(2):75-113.